Self-Attention and CTC for Scalable End-to-end Speech Recognition

Julian Salazar

Scientist, Amazon AI
julsal@amazon.com

Background

- Automatic Speech Recognition (ASR):

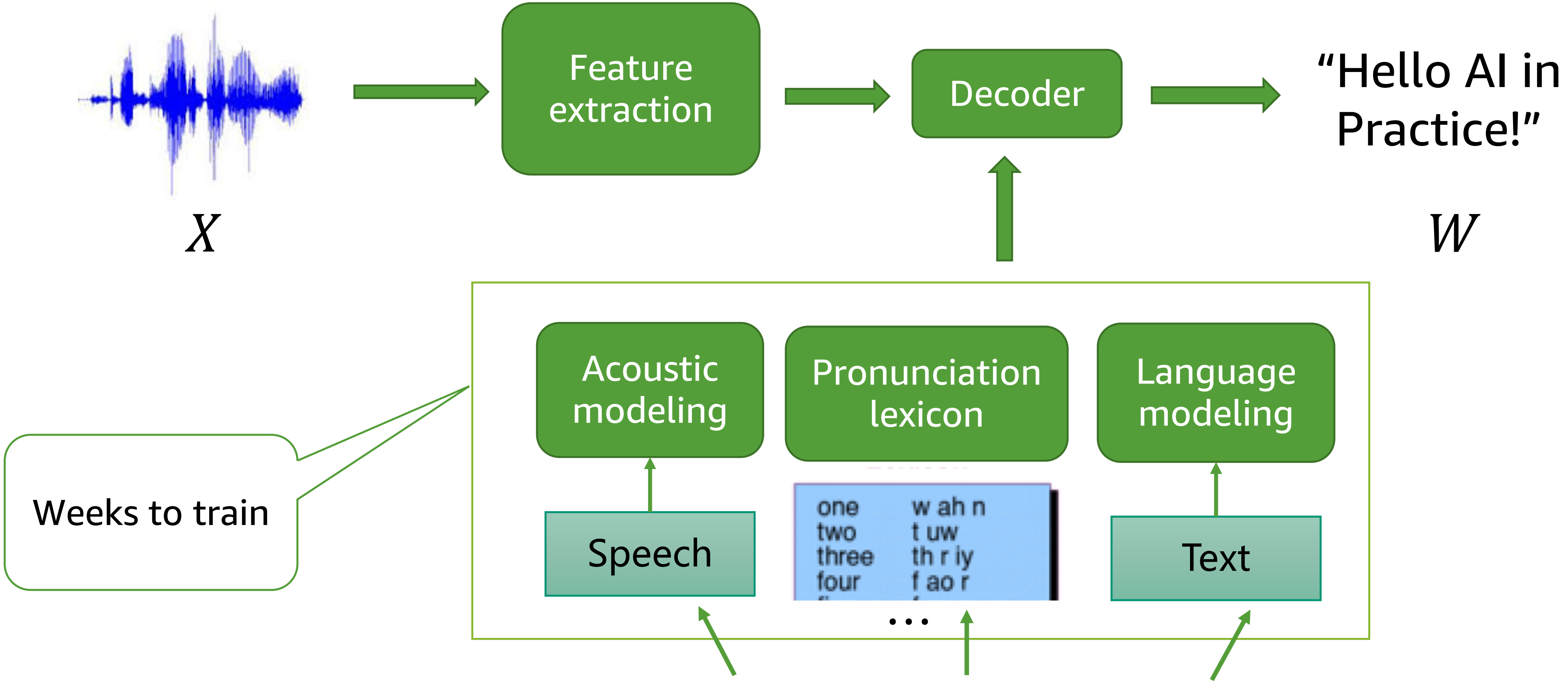
  audio → text

  (+ diarization, punctuation, code-switching, etc.)

- Amazon:
  - As a component: **Echo, Alexa, Lex**
  - As a service: **Transcribe**

amazon Lab126 aws alexa

Classic ASR systems



$$\hat{W} = \mathrm{argmax}_W(p(W|X)) = \mathrm{argmax}_w\left(p(X|Q) * P(Q|W) * P(W)\right)$$

End-to-end ASR systems

X

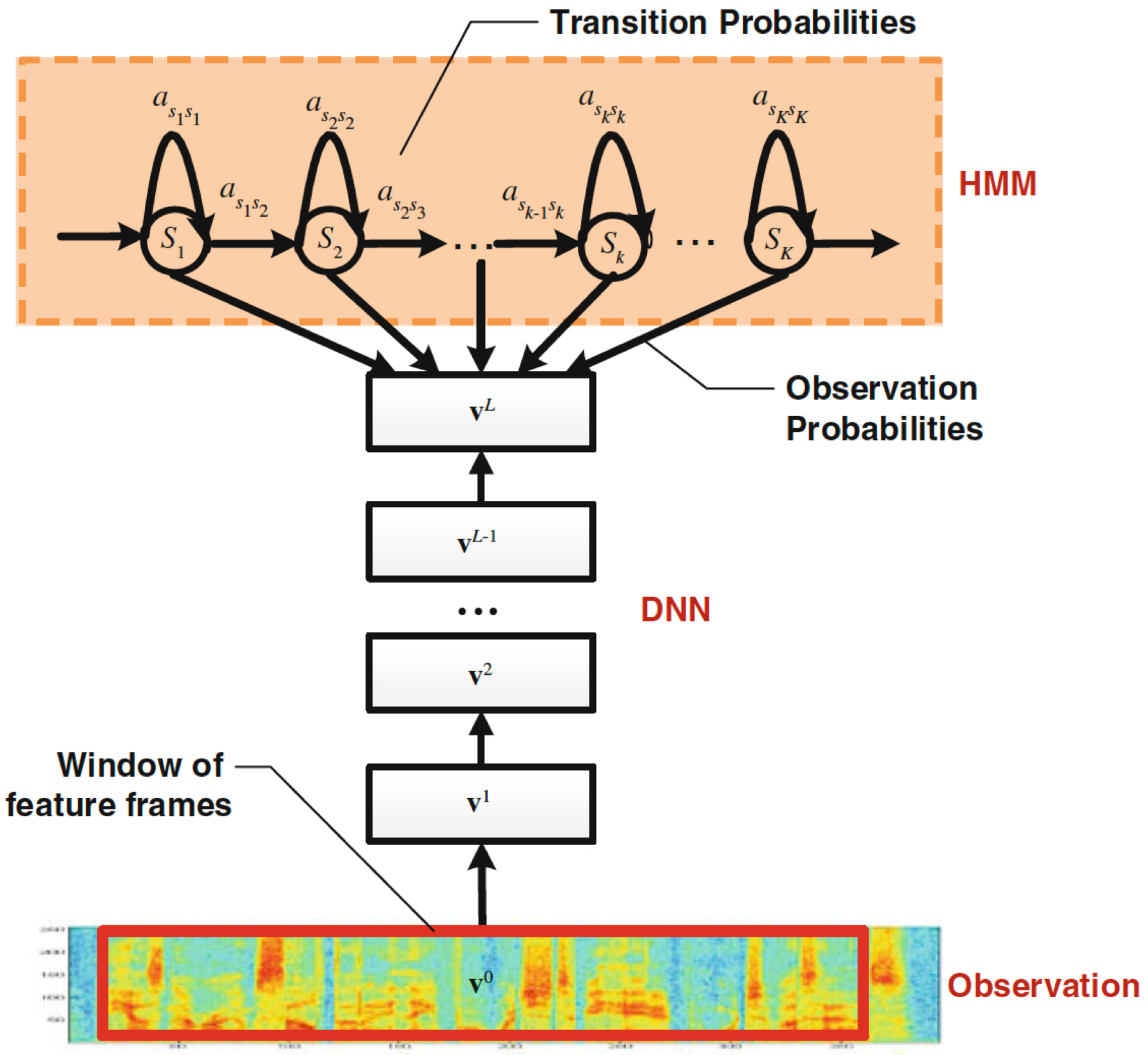
DEEP LEARNING!

$$p(W|X)$$

"Hello AI in Practice!"

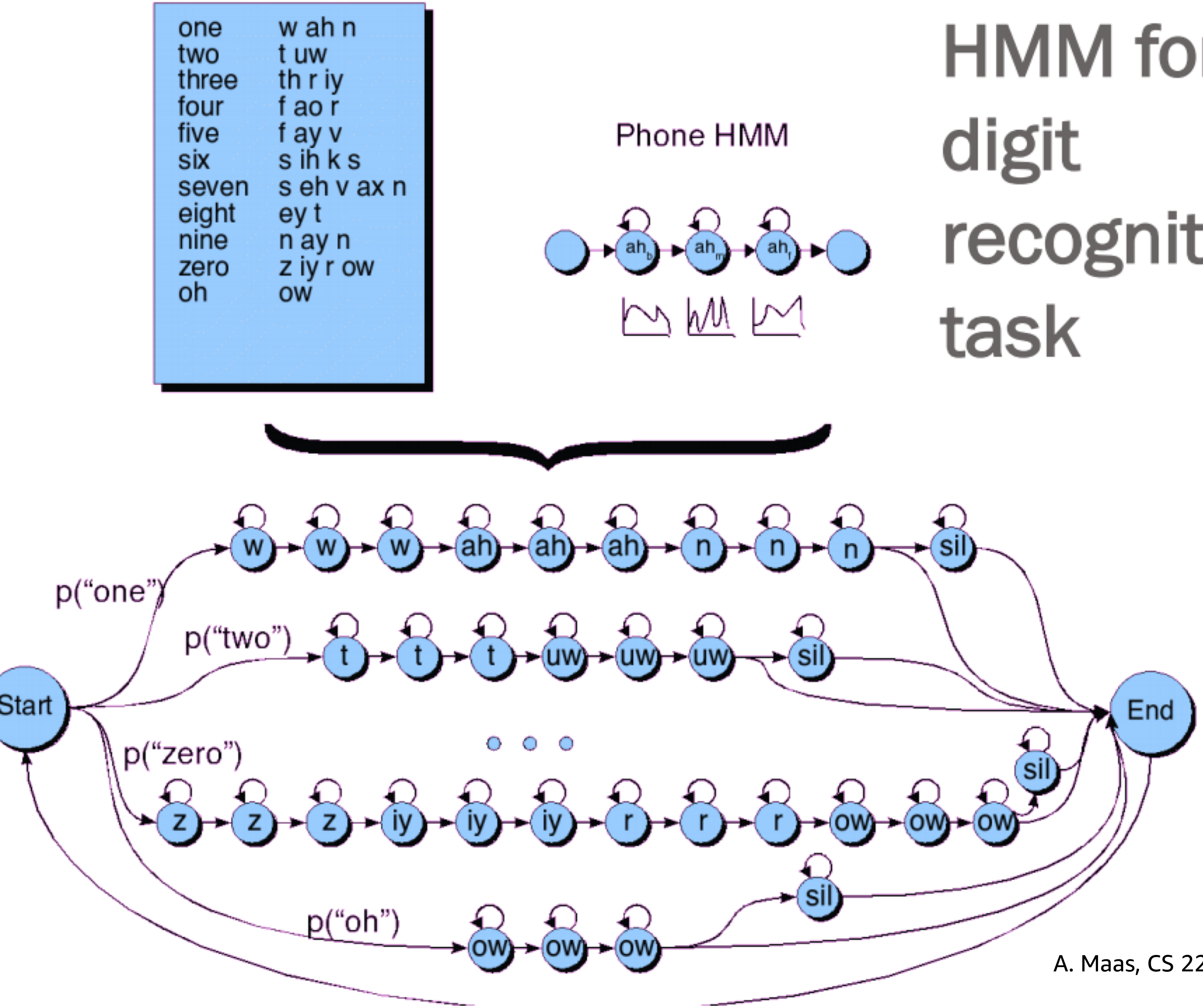
W

ASR: HMM-NN framework



D. Yu and L. Deng, *Automatic Speech Recognition: A Deep Learning Approach*, 2015

HMM for the digit recognition task



A. Maas, CS 224S: Spoken Language Understanding (Stanford), 2017

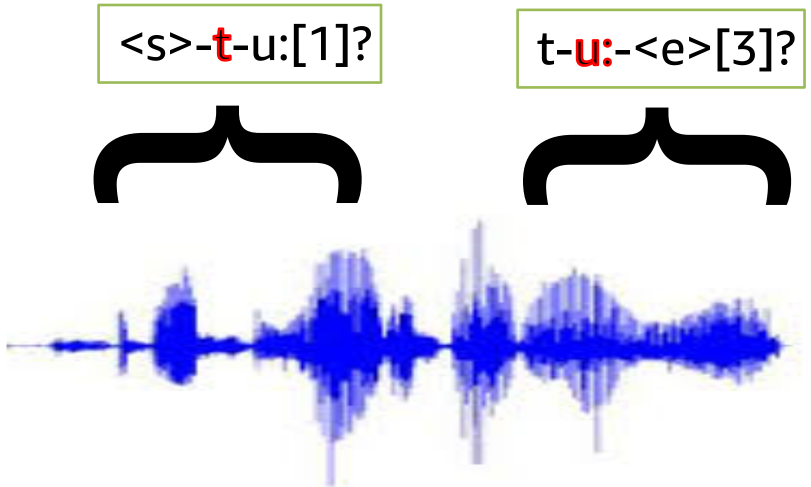
ASR: HMM-NN and alignment

HMM-NN frameworks require forced alignment of training data

Actual training data:

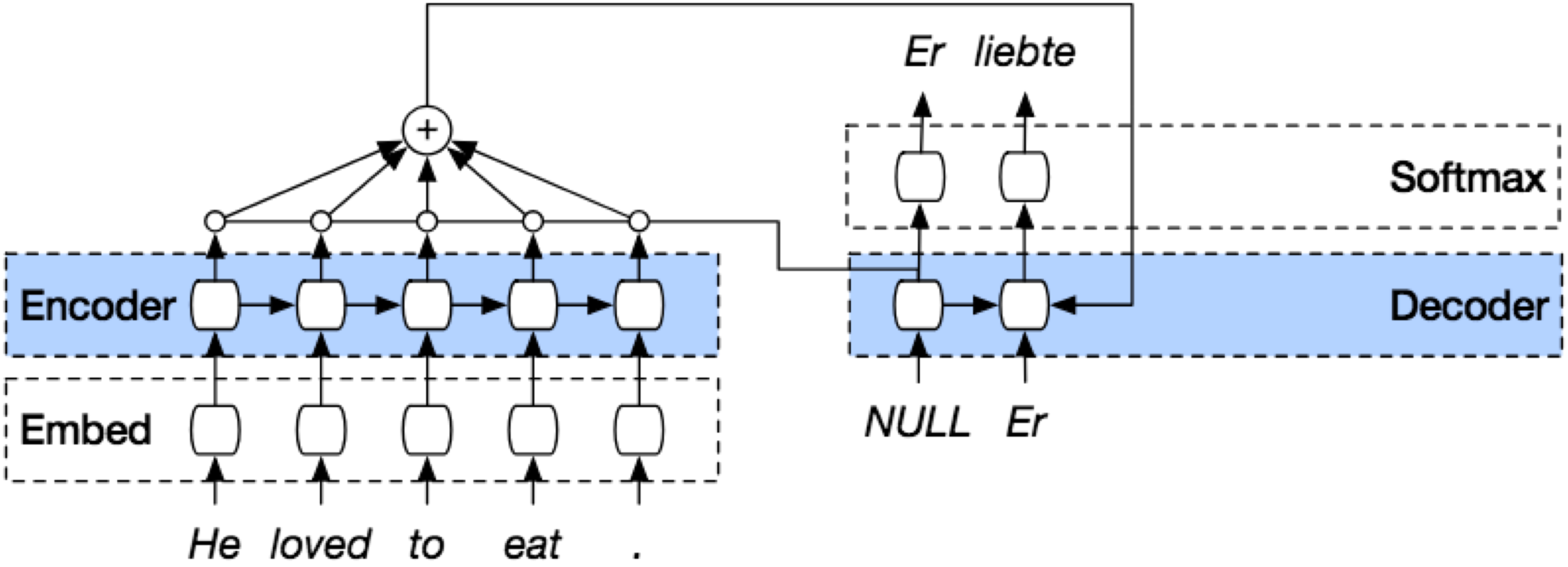
audio file → "to be or not to be"

Need to guess at training time
(e.g., with an existing model!)



ASR: Encoder-decoder (with attention)

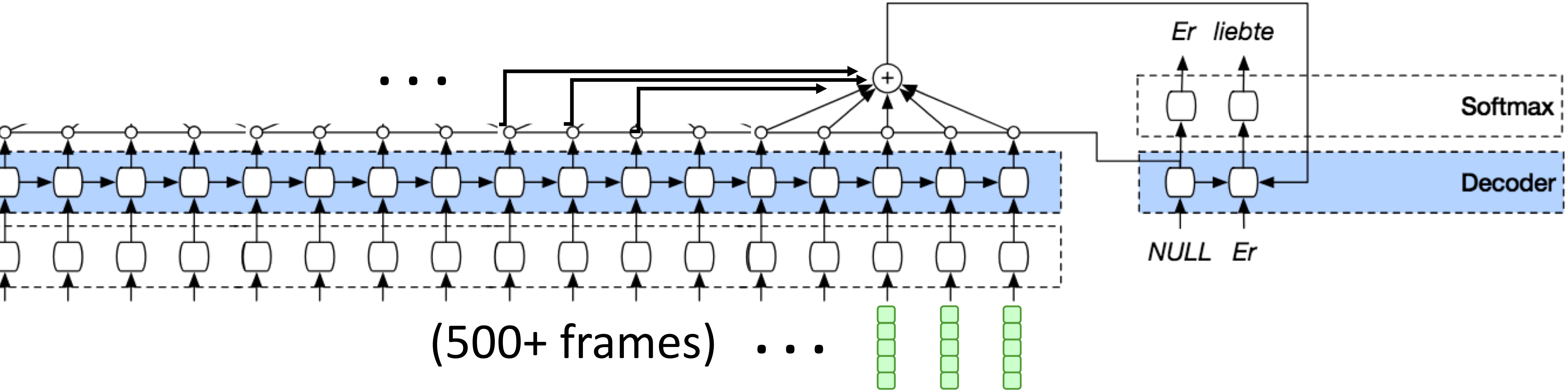
At inference, text is produced “autoregressively”



S. Merity, https://smerity.com/articles/2016/google_nmt_arch.html, 2016

ASR: Encoder-decoder (with attention)

Training on speech is hard!



Self-attention and CTC

Julian Salazar, Katrin Kirchhoff, Zhiheng Huang

"Self-attention networks and connectionist temporal classification for speech recognition"

2019 Proc. of IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP 2019)

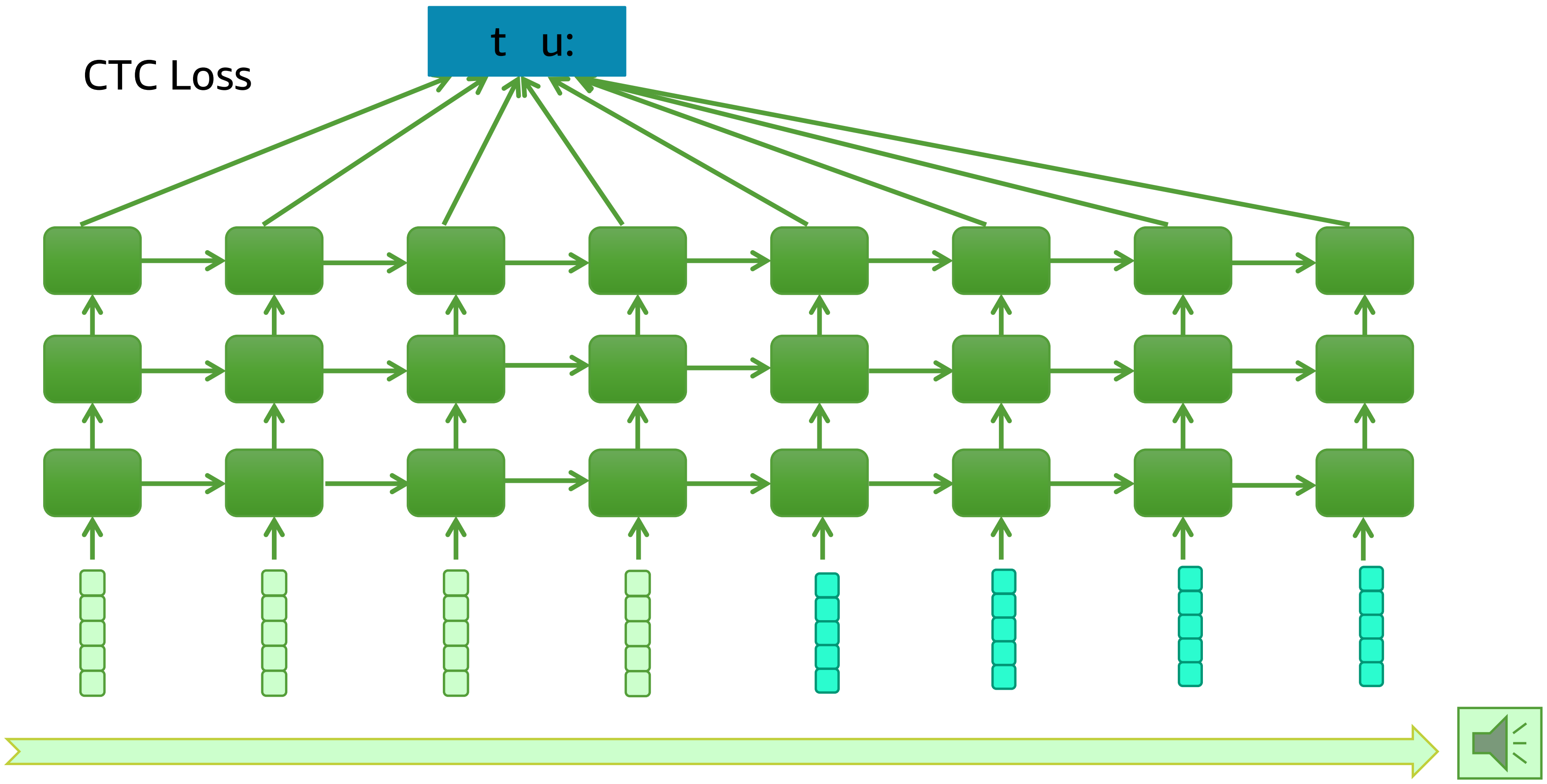
https://arxiv.org/abs/1901.10055

Motivation

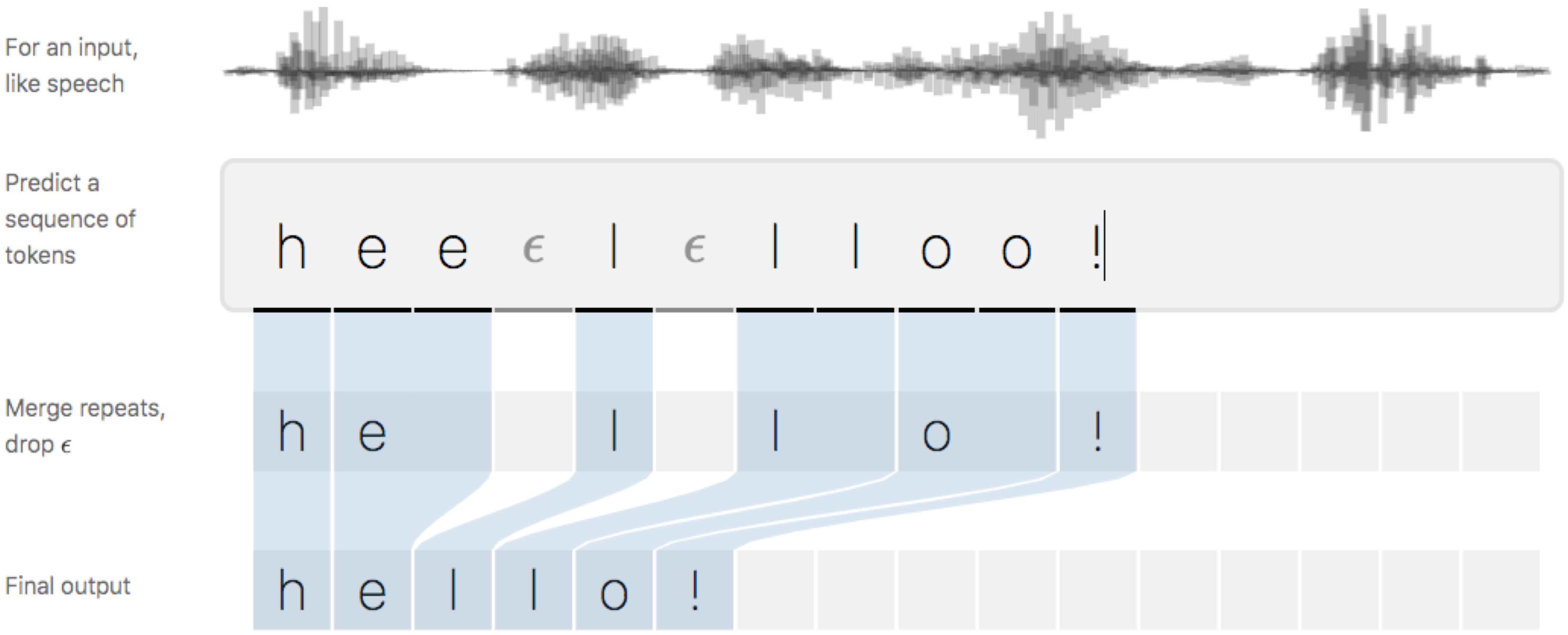
Speech recognition:

- HMM-NN is hand-engineered (too much inductive bias)
- Encoder-decoder is hard to train (too little inductive bias)
- Recurrent models and autoregressive decoding are slow

ASR: CTC framework



Connectionist temporal classification (CTC)

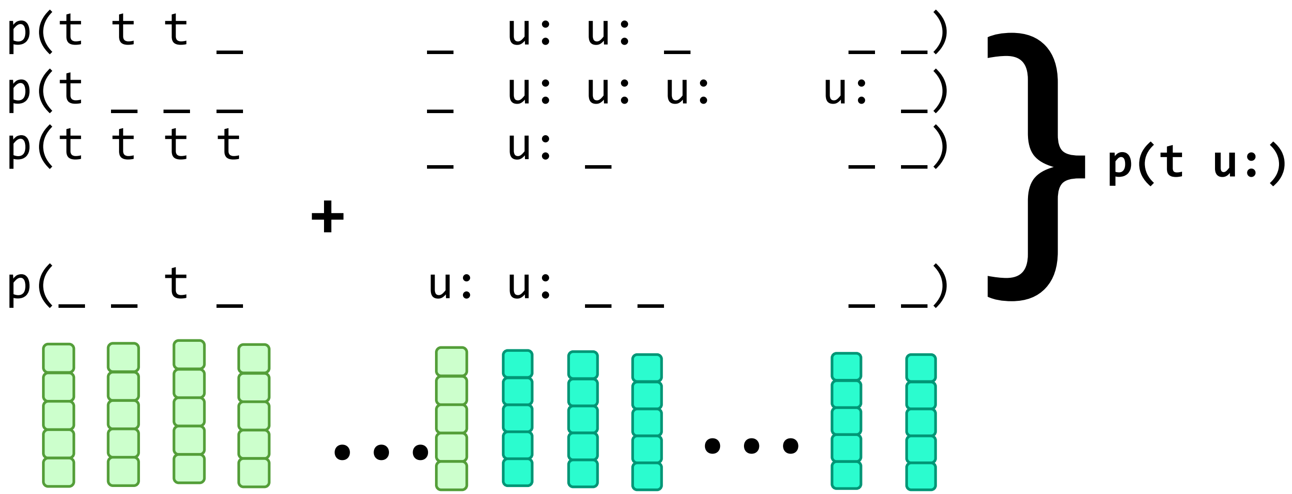


A. Hannun, "Sequence Modeling With CTC", distill.pub 2017

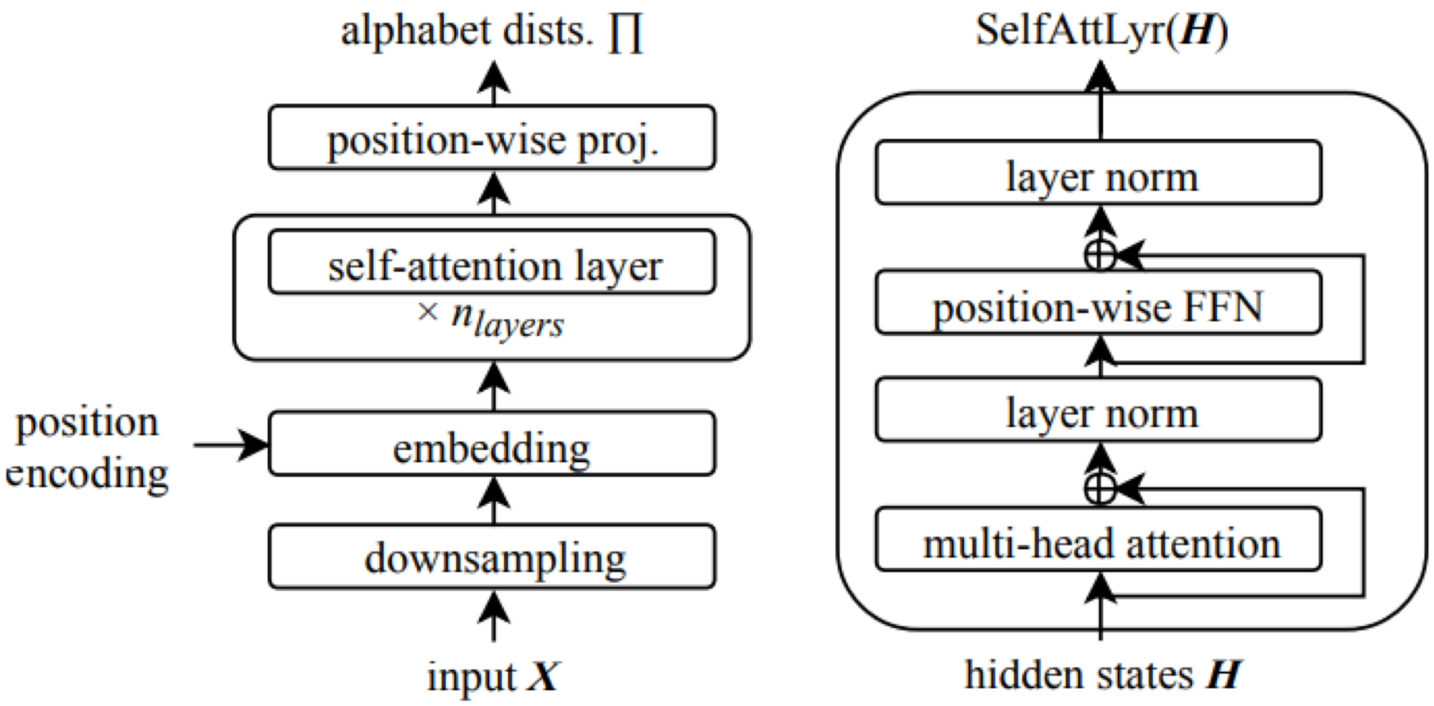
Connectionist temporal classification (CTC)

Inductive biases:

- Monotonicity
- Conditional independence



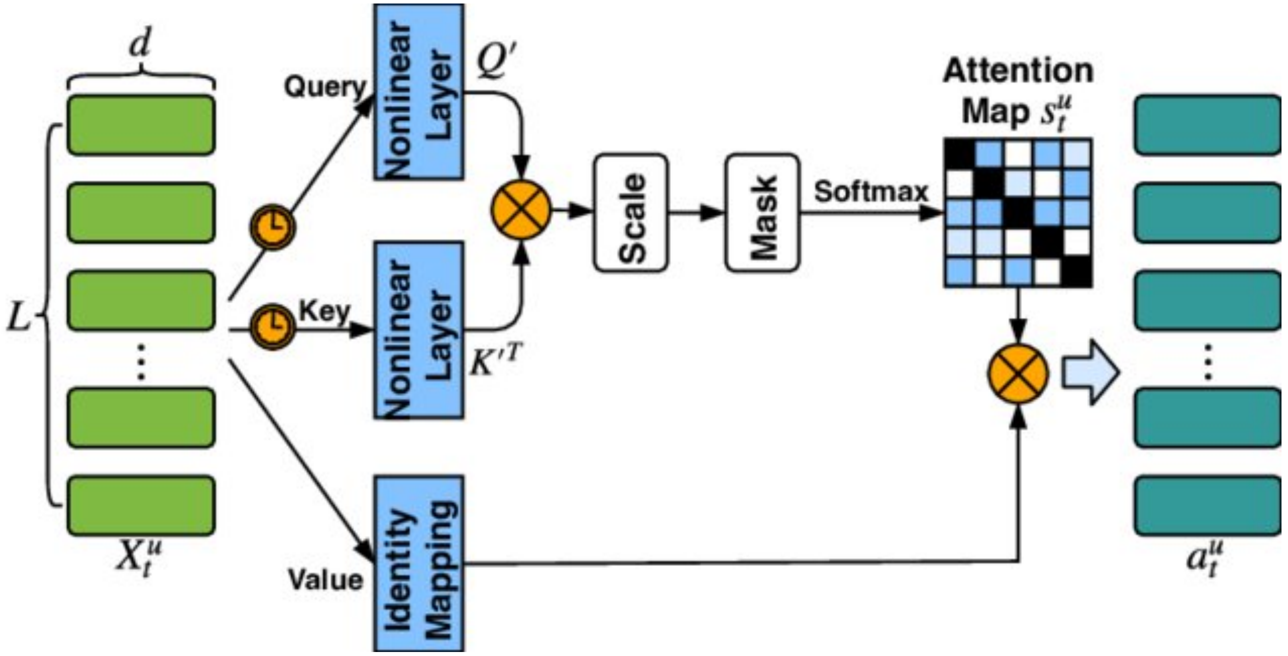
SAN-CTC



(a) SAN-CTC framework (b) Self-attention layer

Fig. 1: Self-attention and CTC

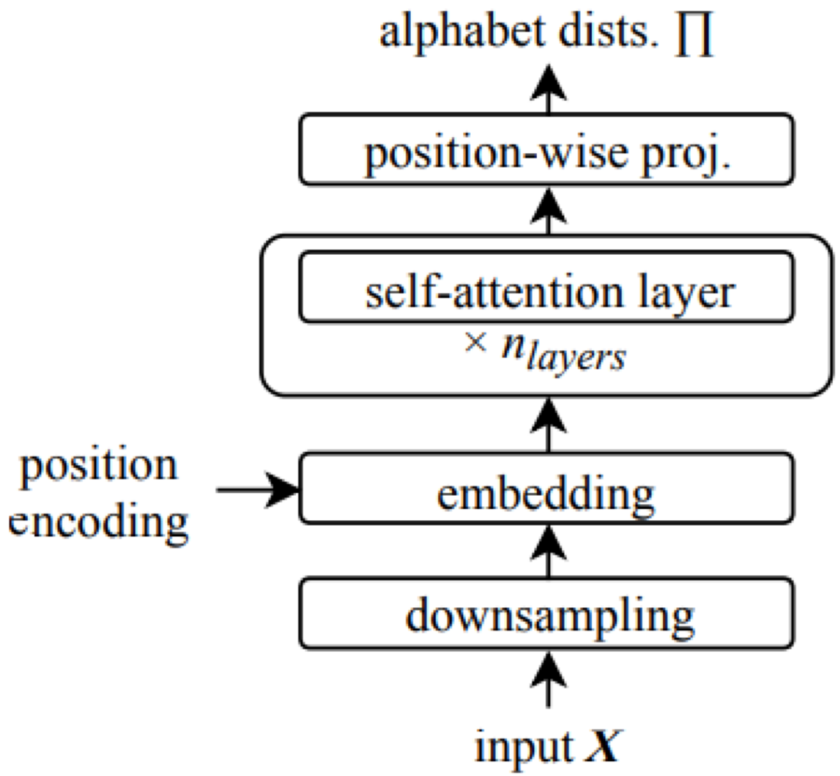
Self-attention



A. Vaswani et. al, "Attention is all you need", NeurIPS 2017
S. Zhang et al, "Next Item Recommendation with Self-Attention", ACM 2018

Featurization

- Choice of alphabet
  - Characters (“t o”, small set)
  - Wordpieces (“to”, larger set)
  - Phonemes (“t u:”, requires dictionary)
- Positional embeddings
- Downsampling/reshaping
  - Self-attention builds an O(T^2) matrix
- Training regimes
  - Inverse-square-root then fixed schedule



WSJ dataset (100 hours)

Character (MLE training, 4-gram LM)
- GatedCNN/Wav2Letter 4.9% char. error → 6.6% word error
- **SAN-CTC: 4.7% char. error → 5.9% word error**
- Encoder-decoder: 3.6% char. error

Phoneme (MLE training, CMU lexicon, 4-gram LM)
- ResCNN-CTC: 5.4% word error
- **SAN-CTC: 5.1% phoneme error → 4.8% word error**
- BRNN/LSTM/CNN-CTC ensemble: → 4.3% word error

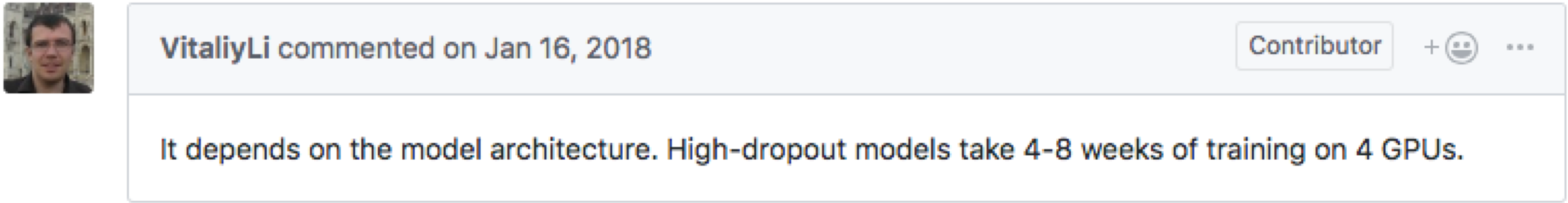
Model	Tok.	test-clean		test-other	
		CER	WER	CER	WER
CTC/ASG (Wav2Letter) [9]	chr.	6.9	7.2	—	—
CTC (DS1-like) [33, 43]	chr.	—	6.5	—	—
Enc-Dec (4-4) [44]	chr.	6.5	—	18.1	—
Enc-Dec (6-1) [45]	chr.	4.5	—	11.6	—
CTC (DS2-like) [8, 32]	chr.	—	5.7	—	15.2
Enc-Dec+CTC (6-1, pretr.) [20]	10k	—	4.8	—	15.3
CTC/ASG (Gated CNN) [23]	chr.	—	4.8	—	14.5
Enc-Dec (2,6-1) [41]	10k	2.9	—	**8.4**	—
CTC (SAN), reshape, additive	chr.	3.2	5.2	9.9	13.9
+ label smoothing, $\lambda = 0.05$	chr.	3.5	5.4	11.3	14.5
CTC (SAN), reshape, concat.	chr.	**2.8**	4.8	9.2	13.1

Table 5: End-to-end, MLE-based, open-vocab. models trained on LibriSpeech. Only WERs incorporating the 4-gram LM are listed.

Performance

Training time (1 Tesla V100):

- 1 week for 70 full passes over LibriSpeech
- Compare w/
  - Transformer Enc-Dec (numbers only on WSJ; comparable)
  - BLSTM Enc-Dec (1 week for 12.5 full passes on GTX 1080Ti)
  - GatedCNN CTC-like [Wav2Letter]:



https://github.com/facebookresearch/wav2letter/issues/11

Performance

Size:

- (10 self-attention layers, 8 heads, 512 hidden dim)
- 30M parameters *(same network for WSJ and LibriSpeech!)*
- Compare w/ 100-250M in Deep Speech 2, Wav2Letter (CTC-like)

Inference time:

- vs. enc-dec: No autoregressive decoding, beam search (much faster)
- vs. BLSTM-CTC [DS2]: 3x+ times faster

AI in practice

Code

POSTED ON DEC 21, 2018 TO AI RESEARCH, ML APPLICATIONS

Open sourcing wav2letter++, the fastest state-of-the-art speech system, and flashlight, an ML library going native

This paper presents a simple end-to-end model for speech recognition, combining a convolutional network based acoustic model and a graph decoding. It is trained to output letters, with transcribed speech, without the need for force alignment of phonemes. We introduce an automatic segmentation criterion for training from sequence annotation without alignment that is on par with CTC [6] while being simpler. We show competitive results in word error rate on the Librispeech corpus

https://code.fb.com/ai-research/wav2letter/
R. Collobert et. al,"Wav2Letter: an End-to-End ConvNet-based Speech Recognition System", arXiv 2016

AI in practice

An All-Neural On-Device Speech Recognizer

Tuesday, March 12, 2019

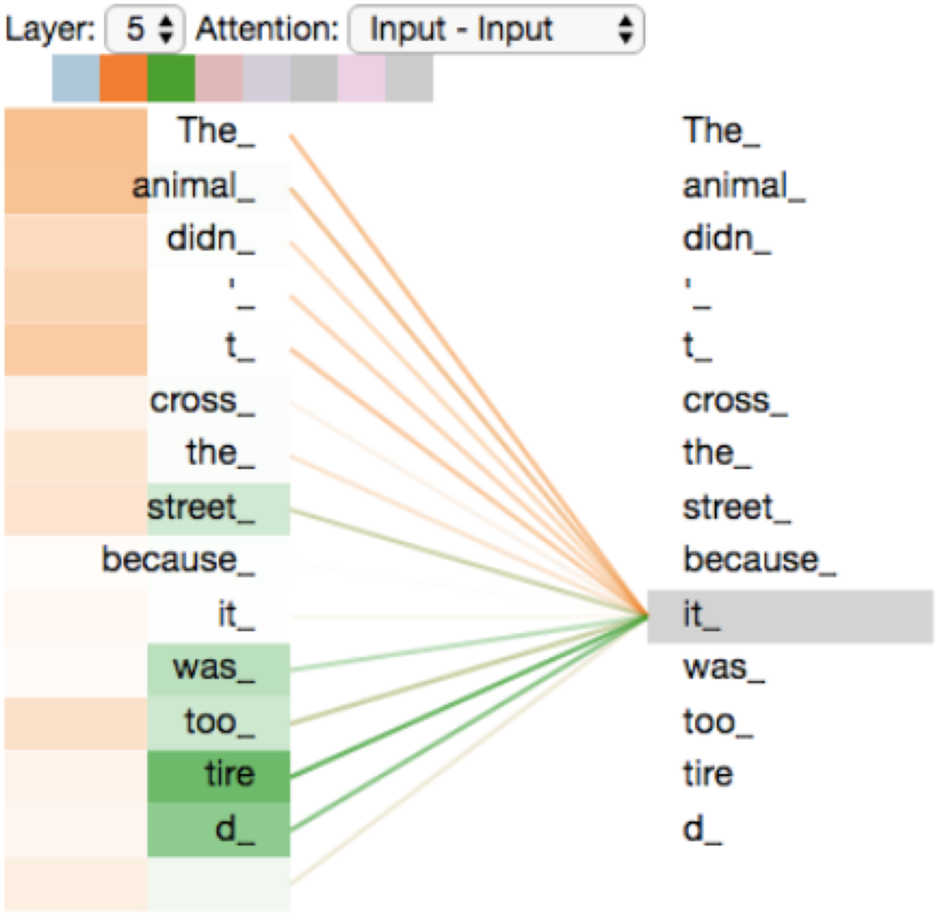
Posted by Johan Schalkwyk, Google Fellow, Speech Team

. . .

Meanwhile, an independent technique called connectionist temporal classification (CTC) had helped halve the latency of the production recognizer at that time. This proved to be an important step in creating the RNN-T architecture adopted in this latest release, which can be seen as a generalization of CTC.

https://ai.googleblog.com/2019/03/an-all-neural-on-device-speech.html
Y. He et. al, “Streaming End-to-end Speech Recognition For Mobile Devices”, ICASSP 2019

Interpreting self-attention



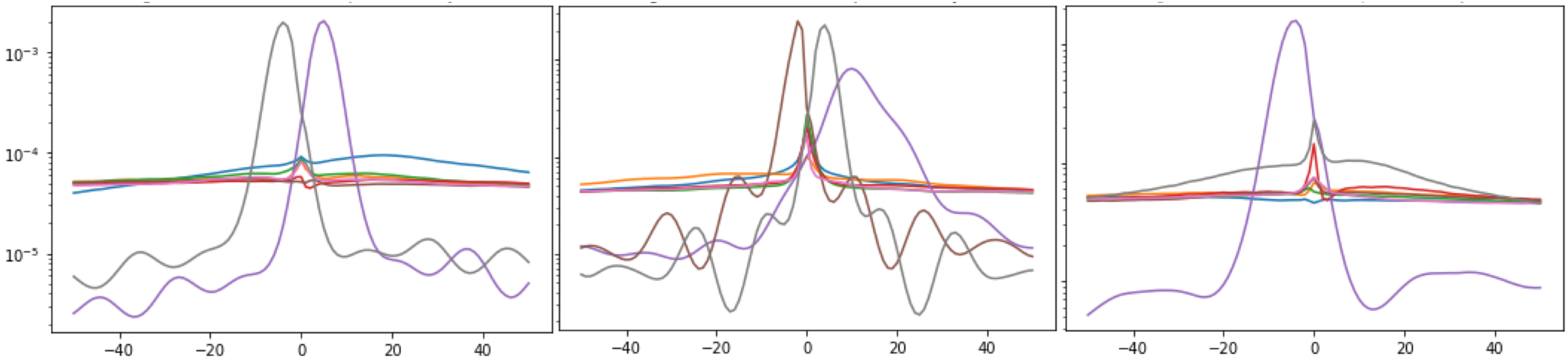
As we encode the word "it", one attention head is focusing most on "the animal", while another is focusing on "tired" -- in a sense, the model's representation of the word "it" bakes in some of the representation of both "animal" and "tired".

J. Alammar, https://jalammar.github.io/illustrated-transformer/, 2018

Interpreting self-attention

- Looka-”head”s for character/wordpiece (spelling, pauses)
- Phonemes are more conditionally independent, so less important

Character: g o o d [sp.] m o r n i n g **Wordpiece:** good morn## ing **CI-phonemes:** g ʊ d [sil.] ˈm ɔː n ɪ ŋ



Interpreting self-attention

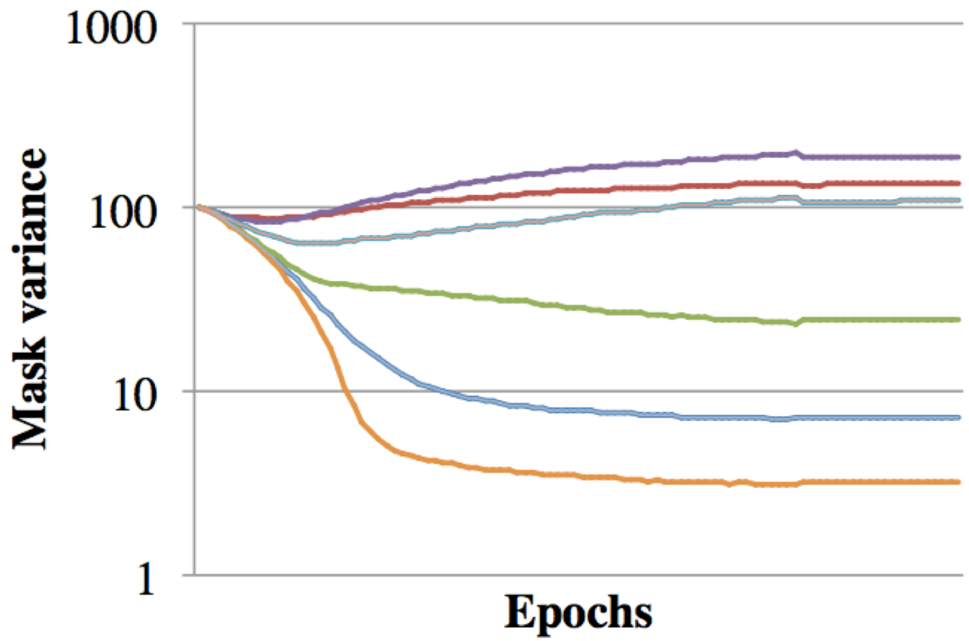


Table 4: *Analysis of function of attention heads. Note that we conducted a small amount of cherry picking by removing 4 outliers that did not seem to fit categories (OY from head 1, ZH from head 3, EH and ER from head 7). Entropy is computed over the correlation scores, truncated below 0.*

i	top phonemes	entropy	comments
1	S, TH, Z	3.7	sibilants
2	</s>	1.9	silence
3	UW, Y, IY, IX	3.6	”you” diphthong
	B, G, D		voiced plosives
	M, NG, N		nasals
4	XM, AW, AA, AY, L, AO, AH	3.2	A, schwa
5	ZH, AXR, R	3.5	R, ZH
6	ZH, Z, S	3.2	sibilants
	IY, IH, Y, UW		”you” diphthong
7	S, </s>, TH CH, SH, F	3.4	fricative, noise
8	mixed	3.7	unfocused

M. Sperber et. al, “Self-attentional acoustic models”, INTERSPEECH 2018

Next steps

- Directed and/or restricted self-attention
- Improved analyses of attention heads
- Learning from tradeoffs between HMM, CTC, seq2seq

Thank you!

J. Salazar, K. Kirchhoff, Z. Huang, *"Self-attention networks and connectionist temporal classification for speech recognition"*, ICASSP 2019

https://arxiv.org/abs/1901.10055

julsal@amazon.com • JulianSlzr.com • @JulianSlzr